Title

Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination.

Journal

The British journal of mathematical and statistical psychology, 73 Suppl 1(S1)

ISSN

0007-1102

Author

Bonett, Douglas G

Publication Date

2020-11-01

DOI

10.1111/bmsp.12189

Peer reviewed

British Journal of Mathematical and Statistical Psychology (2019)

The British Psychological Society

www.wileyonlinelibrary.com

Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination

Douglas G. Bonett*

Department of Psychology, University of California, Santa Cruz, California, USA

The point-biserial correlation is a commonly used measure of effect size in two-group designs. New estimators of point-biserial correlation are derived from different forms of a standardized mean difference. Point-biserial correlations are defined for designs with either fixed or random group sample sizes and can accommodate unequal variances. Confidence intervals and standard errors for the point-biserial correlation estimators are derived from the sampling distributions for pooled-variance and separate-variance versions of a standardized mean difference. The proposed point-biserial confidence intervals can be used to conduct directional two-sided tests, equivalence tests, directional non-equivalence tests, and non-inferiority tests. A confidence interval for an average point-biserial correlation in meta-analysis applications performs substantially better than the currently used methods. Sample size formulas for estimating a point-biserial correlation with desired precision and testing a point-biserial correlation with desired power are proposed. R functions are provided that can be used to compute the proposed confidence intervals and sample size formulas.

1. Introduction

The independent-samples *t*-test is widely used in psychological research to compare two population means that have been estimated from two independent samples. It is common practice to report the results of an independent-samples *t*-test as 'significant' or 'non-significant'. Of course, a 'significant' result does not imply that any practical or scientifically important difference in population means has been detected, and a 'non-significant' result should not be interpreted as evidence of a true null hypothesis. The latest edition of the *Publication Manual of the American Psychological Association* states that 'effect sizes and confidence intervals are the minimum expectations for all APA journals' (American Psychological Association, 2010, p. 33). Point and interval estimates of Cohen's *d* or the point-biserial correlation are recommended supplements to an independent-samples *t*-test (see, for example, Kline, 2013). The point-biserial correlation also is used in psychometric item analyses to assess the association between an item-deleted total score and a dichotomous item score (Crocker & Algina, 1986; Lord & Novick, 1968).

Some researchers prefer the point-biserial correlation as a measure of effect size in a two-group design over Cohen's *d* because of its familiar correlation metric (McGrath

*Correspondence should be addressed to Douglas G. Bonett, Department of Psychology, University of California, Santa Cruz, CA 95064, USA (email: dgbonett@ucsc.edu).

DOI:10.1111/bmsp.12189

& Meyer, 2006). Cohen's *d* is also difficult to interpret if the response variable is not normally distributed (Bonett, 2009). McGrath and Meyer (2006) provide a detailed comparison of Cohen's *d* and the classical point-biserial correlation and conclude that neither measure is universally superior. A neutral stance regarding a preference for Cohen's *d* or the point-biserial correlation is taken here. The purpose of this paper is to present alternative measures of point-biserial correlation, develop a variety of confidence intervals for point-biserial correlations, explain how the proposed confidence intervals can be used to test different types of hypotheses regarding point-biserial correlations, and develop sample size formulas that can be used to design a study to estimate a point-biserial correlation with desired precision or conduct a point-biserial hypothesis test with desired power.

2. Three types of standardized mean differences

A point-biserial correlation can be defined in terms of a standardized mean difference. Three types of standardized mean differences are described here. One type is appropriate for both experimental and non-experimental designs if equal population variances can be assumed; a second type is appropriate for non-experimental designs and does not assume equal population variances; and a third type is appropriate for experimental designs and does not assume equal population variances. In an experimental design with two treatments, a simple random sample is obtained from some population of size N and the sample is then randomly divided into two groups (not necessarily of equal sizes) where one group receives treatment A and the other group receives treatment B. In an experimental design, the sample sizes in each treatment condition are assumed to be fixed. In a non-experimental design, two types of sampling methods are common. With simple random sampling, a random sample is obtained from some population of size N and the members of the random sample are classified into two groups on the basis of some existing characteristic (e.g., male or female). With simple random sampling in a non-experimental design, the total sample size is fixed and the group sample sizes are random. With stratified random sampling, the population of size N is stratified into two subpopulations of sizes N_1 and N_2 on the basis of some existing characteristic. A random sample is then taken from each of the two subpopulations. With stratified random sampling the two sample sizes need not be equal and are assumed to be fixed.

Let X be a dichotomous variable with values 1 and 2. The values of X represent the two treatment conditions in an experimental design or the two subpopulations in a non-experimental design. Let Y denote a quantitative response variable. The most common type of standardized mean difference is defined as

$$\delta_1 = \frac{(\mu_1 - \mu_2)}{\sigma}, \tag{1}$$

where μ_j is the population mean of Y at $X = j$ and σ is an assumed common standard deviation of Y at each level of X. If the homoscedasticity assumption can be justified, δ_1 is appropriate for both experimental and non-experimental designs.

Now consider a two-group non-experimental design where π is the proportion of members in the population that belong to subpopulation 1 and $1 - \pi$ is the proportion of members in the population that belong to subpopulation 2. The variance of Y for all members of the population can be decomposed into between-group and within-group

variances where the within-group variance is $\pi\sigma_1^2 + (1-\pi)\sigma_2^2$. This suggests the standardized mean difference

$$\delta_2 = \frac{(\mu_1 - \mu_2)}{\sqrt{\pi\sigma_1^2 + (1-\pi)\sigma_2^2}}, \tag{2}$$

which is appropriate for non-experimental designs and does not assume homoscedasticity. Note that if $\sigma_1^2 = \sigma_2^2$ then δ_2 reduces to δ_1.

A third type of standardized mean difference that does not assume homoscedasticity is defined as

$$\delta_3 = \frac{(\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}, \tag{3}$$

and is appropriate for experimental designs where participants are randomly assigned to treatment conditions. In a two-group experiment, μ_j and σ_j^2 are respectively the mean and variance of Y, assuming all members of the population had received treatment j. Given that σ_1^2 and σ_2^2 describe the same population in an experimental design, the unweighted average of these two variances is an appropriate description of the average within-treatment variance. Note that if $\sigma_1^2 = \sigma_2^2$ then δ_3 reduces to δ_1.

Estimators of δ_i ($i = 1, 2, 3$) will be used to define different estimators of point-biserial correlation. A frequently used estimator of δ_1, sometimes referred to as Cohen's d, is

$$\hat{\delta}_1 = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\sigma}_p}, \tag{4}$$

where $\hat{\sigma}_p = \sqrt{[(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2]/df}$, n_j is the sample size at $X = j$, $df = n_1 + n_2 - 2$, $\hat{\sigma}_j^2$ is the unbiased estimator of the population variance of Y at $X = j$, and $\hat{\mu}_j$ is the sample mean of Y at $X = j$. The estimator $\hat{\sigma}_p^2$ has two uses. It is an optimal estimator of a common variance and also a consistent estimator of $\pi\sigma_1^2 + (1-\pi)\sigma_2^2$ in non-experimental designs when simple random sampling is used.

The following estimator of δ_2 is appropriate in non-experimental designs with stratified random sampling where π is known:

$$\hat{\delta}_2 = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\pi\hat{\sigma}_1^2 + (1-\pi)\hat{\sigma}_2^2}}, \tag{5}$$

and does not assume homoscedasticity. An estimator of δ_3, which is appropriate in experimental designs and does not assume homoscedasticity (Bonett, 2009), is

$$\hat{\delta}_3 = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}}. \tag{6}$$

3. Alternative measures of point-biserial correlation

If homoscedasticity can be assumed, one type of population point-biserial correlation for non-experimental designs can be defined in terms of δ_1 as

$$\rho_1 = \frac{\delta_1}{\sqrt{\delta_1^2 + \frac{1}{\pi(1-\pi)}}}, \tag{7}$$

where π is the proportion of members in the population that belong to subpopulation 1. If homoscedasticity can be assumed, a second type of population point-biserial correlation for experimental designs can be defined in terms of δ_1 as.

$$\rho_2 = \frac{\delta_1}{\sqrt{\delta_1^2 + 4}}, \tag{8}$$

where π is set to 1/2 to reflect the fact that the entire population could hypothetically be assessed under one treatment condition and the *same* population could also be assessed under a second treatment condition.

A third type of population point-biserial correlation, which does not assume homoscedasticity, can be defined in terms of δ_2 as

$$\rho_3 = \frac{\delta_2}{\sqrt{\delta_2^2 + \frac{1}{\pi(1-\pi)}}}, \tag{9}$$

and is appropriate for non-experimental designs. A fourth type of population point-biserial correlation, which does not assume homoscedasticity, can be defined in terms of δ_2 as

$$\rho_4 = \frac{\delta_3}{\sqrt{\delta_3^2 + 4}}, \tag{10}$$

and is appropriate for experimental designs. Note that ρ_3 is a heteroscedastic alternative to ρ_1 and ρ_4 is a heteroscedastic alternative to ρ_2. With homoscedasticity, $\rho_1 = \rho_3$ and $\rho_2 = \rho_4$.

If the dichotomous variable X is a nominal scale measurement of some attribute (e.g., male versus female or treatment A versus treatment B), the sign of the point-biserial correlation depends on how the two groups are coded and only the absolute magnitude of the point-biserial provides useful information. However, if X represents an ordinal scale measurement of some attribute (e.g., control versus treatment or 1 week of treatment versus 3 weeks of treatment), then both the sign and the magnitude of the point-biserial correlation provide useful information.

In a single-factor between-subjects design, another popular measure of effect size is $\eta^2 = 1 - \sigma^2/\sigma_Y^2$, where σ^2 is an assumed common variance of Y within each level of the between-subjects factor, and σ_Y^2 is the variance of Y. η^2 describes the proportion of the response-variable variance that is predictable from the independent variable. In a two-group non-experimental design, σ_Y^2 is the variance of Y for all members in the population and η^2 can be defined as ρ_1^2. In a two-group experimental design, $\sigma_Y^2 = \sigma^2 + (\mu_1 - \mu_2)^2/4$ (McGrath & Meyer, 2006). In an experimental design η^2 can be defined as ρ_2^2.

Some references claim that a point-biserial correlation has a range from –1 to 1 (Stuart, Ord, & Arnold, 1999, p. 496) while other references (see, for example, Lord & Novick,

1968, p. 340) claim that the point-biserial correlation has a maximum of about .8. Given that δ_i is unbounded, it is clear that ρ_i has a range of –1 to 1. Although ρ_i has a theoretical range of –1 to 1, the values of ρ_1 and ρ_3 depend on the values of π. Table 1 gives the values of ρ_1 corresponding to different values of δ_1 for $\pi = .1$, .3, and .5. The entries in Table 1 suggest that a 'large' point-biserial correlation is smaller than what might be considered to be a 'large' Pearson correlation between two quantitative variables.

The claim that a point-biserial correlation has a maximum value of about .8 applies to the case where Y and X are bivariate normal and X has been artificially dichotomized. In this situation, it can be shown (Gradstein, 1986) that the classical point-biserial correlation has a maximum value of about .8 and that this maximum is achieved at $\pi = 1/2$. If X has been artificially dichotomized, then a biserial correlation is usually a more appropriate measure of association than a point-biserial correlation (Stuart *et al.*, 1999, p. 492).

4. Independence and mean-independence

For two quantitative variables X and Y, a Pearson correlation equal to 0 implies independence of X and Y only in the special case of bivariate normality. For the point-biserial correlations defined here, what does $\rho_i = 0$ ($i = 1, \ldots, 4$) imply? To answer this question, let Y and X represent two random variables and let r and s be positive integers. Goldberger (1991) shows that Y is independent of X if $E(X^r Y^s) = E(X^r)E(Y^s)$ for all r and s, Y is mean-independent of X if $E(X^r Y) = E(X^r)\,E(Y)$ for all r, and Y and X are uncorrelated if $E(XY) = E(X)E(Y)$, assuming these expectations exist. Goldberger also shows that mean-independence implies $E(Y|X) = E(Y)$ for all X. Consequently, if $\mu_1 = \mu_2$ then Y is mean-independent of X. Hence $\rho_i = 0$ ($i = 1, \ldots, 4$) implies that Y is mean-independent of X for any distribution of Y that has a finite mean and variance. Furthermore, $\rho_i = 0$ ($i = 1, \ldots, 4$) implies that Y is independent of X for any distribution of Y if the shape of the distribution of Y is the same at each level of X.

5. Point-biserial estimators

The estimators of a population standardized mean difference given in Section 2 can be used to define estimators of ρ_i ($i = 1, \ldots, 4$). Pearson's classical estimator of ρ_1 can be expressed as (Hays, 1988, p. 311)

Table 1. Values of ρ_1 as a function of δ_1 and π

δ_1	$\pi = .5$	$\pi = .3$	$\pi = .1$
0	0	0	0
0.2	.10	.09	.06
0.5	.24	.22	.15
1.0	.45	.42	.29
1.5	.60	.57	.41
2.0	.71	.68	.51
2.5	.78	.75	.60
3.0	.83	.81	.67

Note. $\rho_1 = \rho_3$ with homoscedasticity, $\rho_1 = \rho_2$ with $\pi = .5$, and $\rho_1 = \rho_4$ with homoscedasticity and $\pi = .5$.

$$\hat{\rho}_1 = \frac{t}{\sqrt{t^2 + df}}, \tag{11}$$

where $t = (\hat{\mu}_1 - \hat{\mu}_2)/\sqrt{\hat{\sigma}_p^2(1/n_1 + 1/n_2)}$, $df = n_1 + n_2 - 2$, and $\hat{\sigma}_p^2$ is the pooled variance estimate used in equation (4). After some algebra, equation (11) can be expressed as

$$\hat{\rho}_1 = \frac{\hat{\delta}_1}{\sqrt{\hat{\delta}_1^2 + \frac{df}{n\hat{\pi}(1-\hat{\pi})}}}, \tag{12}$$

where $n = n_1 + n_2$ and $\hat{\pi} = n_1/n$. In meta-analysis applications, the following estimator of ρ_1 is typically used:

$$\bar{\rho}_1 = \frac{c\hat{\delta}_1}{\sqrt{c^2\hat{\delta}_1^2 + \frac{1}{\hat{\pi}(1-\hat{\pi})}}}, \tag{13}$$

where $c = 1 - 3/(4n - 9)$ is a bias adjustment to $\hat{\delta}_1$ derived by Hedges (1981). The following estimator of ρ_2 is defined using $\hat{\delta}_1$:

$$\hat{\rho}_2 = \frac{\hat{\delta}_1}{\sqrt{\hat{\delta}_1^2 + 4}}. \tag{14}$$

It does not assume homoscedasticity and is appropriate in experimental designs with equal or unequal sample sizes. If π is known and stratified random sampling is used, the following estimator of ρ_3 is defined using $\hat{\delta}_2$:

$$\hat{\rho}_3 = \frac{\hat{\delta}_2}{\sqrt{\hat{\delta}_2^2 + \frac{1}{\pi(1-\pi)}}}. \tag{15}$$

It does not assume homoscedasticity.

The following estimator of ρ_4, which is appropriate for experimental designs with equal or unequal sample sizes and does not assume homoscedasticity, is defined using $\hat{\delta}_3$:

$$\hat{\rho}_4 = \frac{\hat{\delta}_3}{\sqrt{\hat{\delta}_3^2 + 4}}. \tag{16}$$

To summarize, the classical point-biserial estimator ($\hat{\rho}_1$) is appropriate in both experimental and non-experimental designs if homoscedasticity can be assumed. The classical estimator also is appropriate in non-experimental designs with heteroscedasticity if simple random sampling is used so that $\hat{\pi}$ will be a consistent estimator of π. The estimator $\hat{\rho}_2$ is appropriate in experimental designs with equal or unequal sample sizes if homoscedasticity can be assumed, while $\hat{\rho}_4$ is appropriate in experimental designs with equal or unequal sample sizes when homoscedasticity cannot be assumed. The estimator $\hat{\rho}_3$ is appropriate in non-experimental designs with equal or unequal sample sizes when homoscedasticity cannot be assumed and stratified random sampling has been used.

6. Bias of point-biserial correlation estimators

The small-sample bias of $\hat{\rho}_1$ and $\bar{\rho}_1$ is given in Table 2 for a non-experimental design with simple random samples of size $n = 30$ and $n = 60$, three values of π (.20, .35, .50), within-subpopulation normality, and homoscedasticity. The bias was estimated from 200,000 Monte Carlo trials. With simple random sampling, the group sample sizes (n_1 and n_2) are random variables but were constrained to be >2. The classical Pearson estimator ($\hat{\rho}_1$) is nearly unbiased in all conditions, while the alternative estimator ($\bar{\rho}_1$) that is commonly used in meta-analyses can have substantial negative bias.

The small-sample bias of $\hat{\rho}_1$ is given in Table 3 for non-experimental designs with simple random samples of size $n = 30$ and $n = 60$, three values of π (.20, .35, .50), within-subpopulation normality, and heteroscedasticity. As in the homoscedastic case, the small-sample bias of $\hat{\rho}_1$ is negligible under heteroscedasticity with simple random sampling.

The small-sample bias of $\hat{\rho}_2$ and $\hat{\rho}_4$ as estimators of ρ_4 is given in Table 4 for experimental designs where n_1 and n_2 are fixed. Recall that $\rho_2 = \rho_4$ if $\sigma_1/\sigma_2 = 1$. If the sample sizes are equal or if $\sigma_1/\sigma_2 = 1$, the bias of $\hat{\rho}_2$ and $\hat{\rho}_4$ is nearly identical. With $\sigma_1/\sigma_2 = 2$ and unequal sample sizes, $\hat{\rho}_4$ remains nearly unbiased while $\hat{\rho}_2$ (which assumes equal variances) has substantial bias. When the group with the smaller sample size has the larger variance, $\hat{\rho}_2$ has positive bias, and when the group with the larger sample size has the larger variance, $\hat{\rho}_2$ has negative bias. In practice, it can be difficult to determine if the homoscedasticity assumption can be satisfied, and $\hat{\rho}_4$ will be preferred to $\hat{\rho}_2$ in applications where the sample sizes are not equal.

Table 2. Bias of $\hat{\rho}_1$ and $\bar{\rho}_1$ with normality, homoscedasticity, and simple random sampling

		$n = 30$		$n = 60$	
π	ρ_1	bias($\hat{\rho}_1$)	bias($\bar{\rho}_1$)	bias($\hat{\rho}_1$)	bias($\bar{\rho}_1$)
.20	0	.000	.000	.000	.000
	.2	−.002	−.012	−.002	−.007
	.4	−.004	−.023	−.003	−.013
	.6	−.005	−.027	−.004	−.015
	.8	−.004	−.022	−.003	−.012
.35	0	.000	.000	.000	.000
	.2	−.002	−.012	−.001	−.006
	.4	−.003	−.021	−.001	−.011
	.6	−.002	−.024	−.001	−.012
	.8	.000	−.017	.000	−.008
.50	0	.000	.000	.000	.000
	.2	−.001	−.012	−.001	−.006
	.5	−.002	−.021	−.001	−.010
	.6	.000	−.023	.000	−.011
	.8	.002	−.006	.001	−.008

Note. Absolute bias estimates <.001 are reported as .000. The bias estimates were computed from 200,000 Monte Carlo trials. The random sample sizes (n_1 and n_2) were constrained to be >2.

Table 3. Bias of $\hat{\rho}_1$ with normality, heteroscedasticity, and random sample sizes

		$n = 30$		$n = 60$	
π	ρ_1	$\sigma_1/\sigma_2 = 2$	$\sigma_1/\sigma_2 = 4$	$\sigma_1/\sigma_2 = 2$	$\sigma_1/\sigma_2 = 4$
.20	0	.000	.000	.000	.000
	.2	−.002	−.002	−.002	−.002
	.4	−.004	−.004	−.003	−.003
	.6	−.005	−.005	−.004	−.004
	.8	−.004	−.004	−.003	−.003
.35	0	.000	.000	.000	.000
	.2	−.002	−.002	−.001	−.001
	.4	−.003	−.003	−.001	−.001
	.6	−.002	−.002	−.001	−.001
	.8	.000	.000	.000	.000
.50	0	.000	.000	.000	.000
	.2	−.001	−.001	−.001	−.001
	.5	−.002	−.002	−.001	−.001
	.6	.000	.000	.000	.001
	.8	.002	.002	.001	.001

Note. Absolute bias estimates less than .001 are reported as .000. The bias estimates were computed from 200,000 Monte Carlo trials. The random sample sizes (n_1 and n_2) were constrained to be >2.

7. Variance estimates

The variance of $\hat{\rho}_i$ is needed in meta-analysis (Section 8.4) and sample size planning (Section 11). Applying the delta method, the approximate variances of the point-biserial estimators are

$$\widehat{\text{var}}(\hat{\rho}_1) \approx \frac{\left[\frac{df}{n\hat{\pi}(1-\hat{\pi})}\right]^2 \widehat{\text{var}}(\hat{\delta}_1)}{\left[\hat{\delta}_1^2 + \frac{df}{n\hat{\pi}(1-\hat{\pi})}\right]^3}, \tag{17}$$

$$\widehat{\text{var}}(\hat{\rho}_2) \approx \frac{16\widehat{\text{var}}(\hat{\delta}_1)}{\left[\hat{\delta}_1^2 + 4\right]^3}, \tag{18}$$

$$\widehat{\text{var}}(\hat{\rho}_3) \approx \frac{\left[\frac{1}{\pi(1-\pi)}\right]^2 \widehat{\text{var}}(\hat{\delta}_2)}{\left[\hat{\delta}_2^2 + \frac{1}{\pi(1-\pi)}\right]^3}, \tag{19}$$

$$\widehat{\text{var}}(\hat{\rho}_4) \approx \frac{16\widehat{\text{var}}(\hat{\delta}_3)}{\left[\hat{\delta}_3^2 + 4\right]^3}, \tag{20}$$

Table 4. Bias of $\hat{\rho}_2$ and $\hat{\rho}_4$ with normality, homoscedasticity, heteroscedasticity, and fixed sample sizes

			$\sigma_1/\sigma_2 = 1$		$\sigma_1/\sigma_2 = 2$	
n_1	n_2	ρ_4	bias($\hat{\rho}_2$)	bias($\hat{\rho}_4$)	bias($\hat{\rho}_2$)	bias($\hat{\rho}_4$)
15	15	0	.000	.000	.000	.000
		.2	−.004	−.004	−.003	−.003
		.4	−.006	−.006	−.004	−.004
		.6	−.006	−.006	−.004	−.004
		.8	−.003	−.003	−.002	−.002
20	40	0	.000	.000	.000	.000
		.2	−.003	−.002	.020	−.001
		.4	−.004	−.004	.034	−.002
		.6	−.004	−.002	.039	−.002
		.8	−.002	−.001	.029	−.001
40	20	0	.000	.000	.000	.000
		.2	−.003	−.002	−.018	−.002
		.4	−.004	−.004	−.032	−.002
		.6	−.004	−.003	−.037	−.002
		.8	−.001	−.001	−.029	−.001
30	30	0	.000	.000	.000	.000
		.2	−.002	−.002	−.001	−.001
		.5	−.003	−.003	−.002	−.002
		.6	−.003	−.003	−.002	−.002
		.8	−.001	−.001	−.001	−.001

Note. Absolute bias estimates less than .001 are reported as .000. $\rho_2 = \rho_4$ with $\sigma_1/\sigma_2 = 1$. The bias estimates were computed from 200,000 Monte Carlo trials.

where $\widehat{\text{var}}(\hat{\delta}_i)$ is an estimate of the variance of $\hat{\delta}_i$ ($i = 1, 2, 3$). The following approximate variance estimates for $\hat{\delta}_i$ can be derived using an approach described by Bonett (2008a):

$$\widehat{\text{var}}(\hat{\delta}_1) \approx \frac{\hat{\delta}_1^2\left(\frac{1}{df_1}+\frac{1}{df_2}\right)}{8} + \frac{1}{n_1} + \frac{1}{n_2}, \tag{21}$$

$$\widehat{\text{var}}(\hat{\delta}_2) \approx \frac{\hat{\delta}_2^2\left(\frac{1}{df_1}+\frac{1}{df_2}\right)}{8} + \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2 n_1} + \frac{\hat{\sigma}_2^2}{\hat{\sigma}^2 n_2}, \tag{22}$$

$$\widehat{\text{var}}(\hat{\delta}_3) \approx \frac{\hat{\delta}_3^2\left(\frac{\hat{\sigma}_1^4}{df_1}+\frac{\hat{\sigma}_2^4}{df_2}\right)}{8\hat{\sigma}^4} + \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2 df_1} + \frac{\hat{\sigma}_2^2}{\hat{\sigma}^2 df_2}, \tag{23}$$

where $\hat{\sigma}^2 = \pi\hat{\sigma}_1^2 + (1-\pi)\hat{\sigma}_2^2$ for $\widehat{\text{var}}(\hat{\delta}_2)$ and $\hat{\sigma}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$ for $\widehat{\text{var}}(\hat{\delta}_3)$. Equation (23) was given by Bonett (2008a). The above variance estimates assume normality of Y within each level of X.

The variance of Pearson's point-biserial correlation estimator was unknown until Tate (1954) derived the approximation

$$\overline{\text{var}}(\hat{\rho}_1) \approx \frac{(1-\hat{\rho}_1^2)^2\left[1-\frac{3\hat{\rho}_1^2}{2}+\frac{\hat{\rho}_1^2}{4\hat{\pi}(1-\hat{\pi})}\right]}{n}. \tag{24}$$

Both $\overline{\text{var}}(\hat{\rho}_1)$ and $\widehat{\text{var}}(\hat{\rho}_1)$ are approximations that were derived using completely different approaches. It is informative to compare their values with each other and with the true value of $\text{var}(\hat{\rho}_1)$. Table 5 gives the standard deviation of the ρ_1 estimates and the average values of $\sqrt{\overline{\text{var}}(\hat{\rho}_1)}$ and $\sqrt{\widehat{\text{var}}(\hat{\rho}_1)}$ in 200,000 Monte Carlo trials and for a simple random sample of $n = 60$ in each trial. Both $\sqrt{\overline{\text{var}}(\hat{\rho}_1)}$ and $\sqrt{\widehat{\text{var}}(\hat{\rho}_1)}$ tend to slightly understate the true variability of $\hat{\rho}_1$. Both variance estimates appear to be similar in their accuracy and either could be used in meta-analysis or sample size planning applications.

8. Confidence intervals

8.1. Confidence interval for a point-biserial correlation

A confidence interval for a point-biserial correlation can be obtained by first computing a confidence interval for $\delta_i (i = 1, 2, 3)$ and then substituting the lower and upper limits into equations (12), (14), (15) or (16). An approximate $100(1-\alpha)\%$ confidence interval for δ_i $(i = 1, 2, 3)$ is

$$\hat{\delta}_i \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}\left(\hat{\delta}_i\right)} \tag{25}$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution. An alternative to equation (25) for the special case of δ_1 is based on the computationally intensive

Table 5. Comparison of two estimators of $\sqrt{\text{var}(\hat{\rho}_1)}$ with normality, homoscedasticity, and random sample sizes ($n = 60$)

π	ρ_1	$SD(\hat{\rho}_2)$	Average $\sqrt{\widehat{\text{var}}(\hat{\rho}_1)}$	Average $\sqrt{\overline{\text{var}}(\hat{\rho}_1)}$
.20	0	.132	.129	.127
	.2	.126	.124	.123
	.4	.111	.106	.109
	.6	.088	.085	.086
	.8	.052	.047	.051
.35	0	.131	.129	.127
	.2	.124	.123	.121
	.4	.107	.105	.104
	.6	.079	.076	.077
	.8	.042	.040	.041
.50	0	.131	.129	.126
	.2	.124	.123	.121
	.5	.106	.105	.103
	.6	.077	.076	.076
	.8	.042	.040	.040

Note. The estimates were computed from 200,000 Monte Carlo trials. The random sample sizes (n_1 and n_2) were constrained to be greater than 2.

confidence interval for a Student *t* non-centrality parameter (Steiger & Fouladi, 1997) and is recommend if n_1 or n_2 is <10. Equation (25) assumes that the distribution of Y within each level of X is at most moderately non-normal. A bootstrap confidence interval for δ_i (Kelley, 2005) could be used in applications where the data are clearly non-normal and a data transformation (e.g., log, square root, reciprocal) cannot rectify the problem.

Let L and U denote the lower and upper limits of equation (25), respectively. The lower and upper limits of an approximate $100(1-\alpha)\%$ confidence interval for ρ_1 are

$$\left[\frac{L}{\sqrt{L^2+\frac{df}{n\hat{\pi}(1-\hat{\pi})}}}, \frac{U}{\sqrt{U^2+\frac{df}{n\hat{\pi}(1-\hat{\pi})}}}\right], \tag{26}$$

where L and U are the lower and upper limits for δ_1. The lower and upper limits of an approximate $100(1-\alpha)\%$ confidence interval for ρ_3 are

$$\left[\frac{L}{\sqrt{L^2+\frac{1}{\pi(1-\pi)}}}, \frac{U}{\sqrt{U^2+\frac{1}{\pi(1-\pi)}}}\right], \tag{27}$$

where L and U are the lower and upper limits for δ_2. The lower and upper limits of approximate $100(1-\alpha)\%$ confidence intervals for ρ_2 or ρ_4 are

$$\left[\frac{L}{\sqrt{L^2+4}}, \frac{U}{\sqrt{U^2+4}}\right], \tag{28}$$

where L and U are the lower and upper limits for δ_1 or δ_3, respectively. The R function `ci.pbcor124` (Appendix) computes the confidence intervals for ρ_1, ρ_2, and ρ_4. The R function `ci.pbcor3` (Appendix) computes the confidence interval for ρ_3.

8.2. Confidence intervals for ρ_1^2 and ρ_2^2

If the confidence interval for ρ_1 does not include 0, then an approximate confidence interval for ρ_1^2 is computed by simply squaring the endpoints of the confidence interval for ρ_1. If the confidence interval for ρ_1 includes 0, the lower limit of the confidence interval for ρ_1^2 is set to 0 and the upper limit is equal to the larger of the squared lower limit or squared upper limit. The same procedure is used to compute an approximate confidence interval for ρ_2^2.

8.3. Confidence interval for difference in point-biserial correlations

A test for equal Pearson correlations using independent samples is discussed in many statistics texts for psychologists (Cohen, Cohen, West, & Aiken, 2003, p. 49; see Howell, 2007, p. 259; Field, Miles, & Field, 2012, p. 239). A confidence interval for the difference of two point-biserial correlations can be used to test for equal population point-biserial correlations and also provides useful information about the magnitude of the difference. Let ρ_{i1} represent a population point-biserial correlation between X and Y in population 1, and let ρ_{i2} represent a population point-biserial correlation between X and Y in population 2. A random sample from population 1 will be used to estimate ρ_{i1} and a random sample from population 2 will be used to estimate ρ_{i2}. Let $\hat{\rho}_{i1}$ represent a point-

biserial estimator from sample 1, and let $\hat{\rho}_{i2}$ represent a point-biserial estimator from sample 2. Let L_1 and U_1 denote the lower and upper $100(1-\alpha)\%$ confidence interval endpoints for ρ_{i1}, and let L_2 and U_2 denote the lower and upper $100(1-\alpha)\%$ confidence interval endpoints for ρ_{i2}. Applying a method described by Zou (2007), an approximate $100(1-\alpha)\%$ confidence interval for $\rho_{i1}-\rho_{i2}$ is

$$\left[\hat{\rho}_{i1}-\hat{\rho}_{i2}-\sqrt{(\hat{\rho}_{i1}-L_1)^2+(\hat{\rho}_{i2}-U_2)^2},\hat{\rho}_{i1}-\hat{\rho}_{i2}+\sqrt{(\hat{\rho}_{i1}-U_1)^2+(\hat{\rho}_{i2}-L_2)^2}\right]. \quad (29)$$

The R function `ci.diff.pbcor` (Appendix) computes expression (29).

8.4. Confidence interval for an average of point-biserial correlations

If a point-biserial correlation is estimated in $m \geq 2$ different studies using the same X and Y variables, we can estimate $\rho = \sum_{k=1}^{m} \rho_{ik}/m$, where ρ_{ik} is the population point-biserial correlation between X and Y that has been estimated in study k ($k = 1, \ldots, m$). The confidence interval for ρ can be substantially narrower than the confidence interval for ρ_{ik} from any single study. Combining parameter estimates from two or more studies is referred to as a meta-analysis.

An estimate of $\rho = \sum_{k=1}^{m} \rho_{ik}/m$ is

$$\hat{\rho} = \frac{\sum_{k=1}^{m} \hat{\rho}_{ik}}{m}, \quad (30)$$

and its estimated variance is

$$\widehat{\text{var}}(\hat{\rho}) = m^{-2} \sum_{k=1}^{m} \widehat{\text{var}}(\hat{\rho}_{ik}), \quad (31)$$

where $\widehat{\text{var}}(\hat{\rho}_{ik})$ is given by (17), (18), (19) or (20) for each k. Note that $\hat{\rho}$ can be an average of different types of point-biserial estimates. For example, $\hat{\rho}_1$ could be computed for non-experimental design studies that used simple random sampling, $\hat{\rho}_3$ could be computed for non-experimental design studies that used stratified random sampling, and $\hat{\rho}_4$ could be computed for experimental design studies.

Following an approach described by Bonett (2008b) for Pearson correlations, an approximate $100(1-\alpha)\%$ confidence interval for ρ is obtained in two steps. In the first step compute

$$\hat{\rho}^* \pm z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}}(\hat{\rho})}{(1-\hat{\rho}^2)^2}} \quad (32)$$

where $\hat{\rho}* = \frac{1}{2}\ln\left(\left[\frac{1+\hat{\rho}}{1-\hat{\rho}}\right]\right)$ is a Fisher transformation of the average point-biserial correlation. The Fisher transformation of the average point-biserial correlation is not a variance-stabilizing transformation, but the sampling distribution of $\hat{\rho}^*$ will be more closely approximated by a normal distribution than the sampling distribution of $\hat{\rho}$.

Let L^* and U^* denote the endpoints of equation (32). In the second step, reverse-transform the endpoints of equation (32) to obtain the following $100(1-\alpha)\%$ confidence interval for ρ:

$$\left[\frac{\exp(2L^*)-1}{\exp(2L^*)+1}, \frac{\exp(2U^*)-1}{\exp(2U^*)+1}\right]. \tag{33}$$

This confidence interval for ρ is a varying-coefficient confidence interval that does not make the unrealistic assumptions of the traditional constant-coefficient ('fixed-effect') and random-coefficient ('random-effects') meta-analysis methods (see Bonett & Price, 2015). The R function `ci.ave.pbcor` (Appendix) computes expression (33).

Unlike the constant-coefficient estimator, expression (33) does not assume equality of ρ_{ik} values, and unlike the random-coefficient estimator, expression (33) does not assume that the ρ_{ik} parameters are a random sample from a normally distributed superpopulation of point-biserial correlations. The traditional constant-coefficient and random-coefficient methods for point-biserial correlations compute a confidence interval for an average Fisher-transformed point-biserial correlation and then reverse-transform the endpoints. Except in some special cases (e.g., a set of correlations that are symmetrically distributed around 0), reverse-transforming an average of Fisher-transformed correlations will introduce bias into the reverse-transformed estimator. The random-coefficient method computes a weighted average estimator of ρ and the weights are assumed to be uncorrelated with the Fisher-transformed point-biserial correlations. A correlation between the weights and the estimates introduces additional bias into the random-coefficient estimator of ρ (Bonett & Price, 2015).

The varying-coefficient confidence interval describes the average of the m population point-biserial correlations. The allure of the random-coefficient confidence interval is that it describes the average of all point-biserial correlations in the superpopulation. However, the random-coefficient confidence interval enjoys this useful interpretation only if the m population point-biserial correlations are a random sample from some definable superpopulation of point-biserial correlations.

8.5. Subgroup analysis

Subgroup analyses are often performed in a meta-analysis to assess the effect of a categorical moderator variable (see Borenstein, Hedges, Higgins, & Rothstein, 2009, Chapter 19). Expressions (33) and (29) can be used in conjunction to perform a two-level subgroup analysis of point-biserial correlations. A two-level point-biserial subgroup comparison can be expressed as $\rho_{iA} - \rho_{iB}$, where ρ_{iA} is the average or two or more point-biserial correlations and ρ_{iB} is the average or two or more point-biserial correlations where the point-biserial correlations that comprise ρ_{iA} are distinct from the point-biserial correlations that comprise ρ_{iB}. In some subgroup analyses, ρ_{iA} or ρ_{iB} might represent a single point-biserial correlation rather than an average. Some examples of subgroups comparisons are $(\rho_{i1} + \rho_{i2})/2 - \rho_{i3}$ and $(\rho_{i1} + \rho_{i2}) - (\rho_{i3} + \rho_{i4} + \rho_{i5})/3$. To compute a 100 $(1-\alpha)$% confidence interval for $\rho_{iA} - \rho_{iB}$, compute $100(1-\alpha)$% confidence intervals for ρ_{iA} and ρ_{iB} and then plug the point and interval estimates into expression (29).

8.6. Confidence interval interpretation

Classical confidence intervals are typically motivated using a relative frequency definition of probability, and statisticians have correctly pointed out that nothing can be said about a computed confidence interval using a relative frequency argument (Arnold, 1990, p. 568). Although the relative frequency definition is of no use after the data have been analysed (Pearson, 1947), a subjective degree of belief definition of probability can be used to

interpret a computed confidence interval, as explained by Bonett and Wright (2007). It is not correct to assume that a subjective degree of belief definition of probability only can be used in a Bayesian analysis.

9. Performance of confidence intervals

The small-sample coverage probabilities of 95% confidence intervals for ρ_1 and ρ_3 were assessed in a Monte Carlo study to simulate a non-experimental design with simple random sampling. Normality within each subpopulation was assumed. Two values of π were examined (.20 and .50) under both homoscedasticity and heteroscedasticity ($\sigma_1/\sigma_2 = 2$). Recall that the estimators of both ρ_1 and ρ_3 are appropriate with heteroscedasticity if simple random sampling is used. Random data were computer generated for five different values of ρ_1 (0, .2, .4, .6, .8). With homoscedasticity, $\rho_1 = \rho_3$, and with heteroscedasticity $\hat{\rho}_1$ is a consistent estimator of ρ_3 if simple random sampling is used. For each value of π and ρ_3, confidence intervals for ρ_1 and $\rho_3$were computed from 100,000 Monte Carlo trials. The estimated coverage probabilities of the two confidence interval methods are summarized in

Table 6. Estimated coverage probabilities for ρ_1 and ρ_3 with normality and random sample sizes

			$\sigma_1/\sigma_2 = 1$		$\sigma_1/\sigma_2 = 2$	
n	π	ρ_3	CI for ρ_1	CI for ρ_3	CI for ρ_1	CI for ρ_3
30	.20	0	.956	.926	.956	.927
		.2	.954	.931	.955	.930
		.4	.950	.942	.949	.942
		.6	.944	.957	.945	.958
		.8	.943	.973	.942	.974
	.50	0	.948	.947	.949	.947
		.2	.949	.947	.948	.948
		.4	.948	.948	.949	.948
		.6	.950	.950	.950	.949
		.8	.952	.952	.951	.951
60	.20	0	.952	.939	.952	.939
		.2	.950	.941	.951	.941
		.4	.945	.949	.945	.948
		.6	.938	.959	.938	.960
		.8	.928	.973	.928	.973
	.50	0	.949	.949	.949	.949
		.2	.949	.948	.949	.949
		.4	.949	.949	.949	.949
		.6	.950	.950	.950	.950
		.8	.951	.950	.950	.950

Note. $\rho_1 = \rho_3$ with $\sigma_1/\sigma_2 = 1$ and $\hat{\rho}_1$ is a consistent estimator of ρ_3 with $\sigma_1/\sigma_2 \neq 1$ if simple random sampling is used. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. The random sample sizes (n_1 and n_2) were constrained to be >2.

Table 6. Both 95% confidence intervals have estimated coverage probabilities that are close to .95 under all conditions examined.

In non-experimental designs where π is known, stratified random sampling can be used to obtained the desired sample sizes in each group. Researchers often want equal sample sizes to maximize the power of the independent-samples *t*-test or to minimize the negative effects of assumption violations (Scheffé, 1959). A second Monte Carlo study examined the performance of the confidence intervals for ρ_1 and ρ_3 with stratified random sampling from a population with $\pi = .20$. The results are summarized in Table 7. Unless the fixed sample sizes are selected such that $n_1/(n_1 + n_2) = \pi$, the coverage probability of a 95% confidence interval for ρ_1 can be far below .95. Poor performance was observed when $\rho_1 = .8$. But with $\pi = .20$, a ρ_1 value of .8 corresponds to a δ_1 value of about 3.33, which rarely would be observed in any actual study. In non-experimental designs where the minority subpopulation is oversampled in an effort to obtain similar sample sizes, the classical point-biserial correlation should not be used and ρ_3 is the recommended alternative.

The small-sample coverage probabilities of 95% confidence intervals for ρ_2 and ρ_4 (which are appropriate for experimental designs) were estimated from 100,000 Monte Carlo trials for fixed sample sizes, within-condition normality, homoscedasticity, and heteroscedasticity ($\sigma_1/\sigma_2 = 2$). The results are summarized in Table 8. With homoscedasticity, $\rho_2 = \rho_4$, but with heteroscedasticity $\hat{\rho}_2$ is not a consistent estimator of ρ_4 if the sample sizes are unequal. The confidence interval for ρ_2 has coverage probabilities that were close to .95 in the equal sample size conditions under both homoscedasticity and heteroscedasticity. In the heteroscedastic cases with unequal sample sizes, the confidence interval for ρ_2 is liberal when the group with the larger variance has the smaller sample size and is conservative when the group with the larger variance has the larger sample size. The confidence interval for ρ_2 is liberal with moderate and large values of ρ_2, heteroscedasticity, and equal sample sizes. The confidence interval for ρ_4 has coverage probabilities that are close to .95 under all conditions.

Confidence intervals for standardized mean differences are not robust to violations of the normality assumption (Bonett, 2009). The coverage probability for a standardized mean difference tends to be conservative with platykurtic (short-tailed) distributions and anti-conservative with leptokurtic (long-tailed) distributions within each level of X. Confidence intervals for point-biserial correlations will have similar properties. The coverage probabilities of 95% confidence intervals for ρ_1 under non-normality and homoscedasticity assuming a simple random sample of size $n = 60$ were estimated from 100,000 Monte Carlo trials using four different beta distributions. The results are summarized in Table 9. The beta distribution family has a finite range and is a useful representation of the distribution of the numerous finite-range tests and questionnaires used in psychology. The Beta(1, 1) and Beta(2, 2) distributions are symmetric and platykurtic with kurtosis coefficients of 1.8 and 2.14, respectively. The Beta(2, 4) and Beta(1, 5) distributions are skewed with skewness coefficients of 0.47 and 1.18, respectively, and with kurtosis coefficients of 2.62 and 4.2, respectively. The coverage probabilities of 95% confidence intervals for ρ_2 under non-normality and homoscedasticity with fixed sample sizes were estimated from 100,000 Monte Carlo trials using the four different beta distributions described above. The results are summarized in Table 10.

The confidence intervals ρ_j are robust to non-normality with small values of ρ_j. With large values of ρ_j, the coverage probabilities can be unacceptably conservative with platykurtic distributions and unacceptably liberal with leptokurtic distributions. Data

Table 7. Estimated coverage probabilities for ρ_1 and ρ_3 with normality, fixed sample sizes, and $\pi = .2$

			$\sigma_1/\sigma_2 = 1$		$\sigma_1/\sigma_2 = 2$	
ρ_3	n_1	n_2	CI for ρ_1	CI for ρ_3	CI for ρ_1	CI for ρ_3
0	15	15	.948	.948	.948	.948
	20	40	.950	.947	.950	.947
	40	20	.950	.948	.950	.948
	30	30	.949	.949	.949	.949
	12	48	.951	.941	.951	.941
.2	15	15	.938	.947	.938	.947
	20	40	.941	.948	.941	.948
	40	20	.941	.945	.941	.945
	30	30	.931	.949	.931	.948
	12	48	.953	.943	.953	.943
.4	15	15	**.908**	.944	**.908**	.944
	20	40	**.912**	.949	**.912**	.949
	40	20	**.913**	.938	**.912**	.937
	30	30	**.870**	.944	**.870**	.944
	12	48	.958	.950	.958	.950
.6	15	15	**.851**	.937	**.851**	.938
	20	40	**.859**	.951	**.859**	.950
	40	20	**.859**	**.923**	**.859**	**.923**
	30	30	**.757**	.937	**.757**	.937
	12	48	.965	.960	.965	.960
.8	15	15	**.754**	.928	**.756**	.927
	20	40	**.769**	.953	**.768**	.953
	40	20	**.769**	**.898**	**.768**	**.899**
	30	30	**.571**	.926	**.572**	.927
	12	48	**.976**	.974	**.976**	.974

Note. $\hat{\rho}_1$ is not a consistent estimator of ρ_3 unless $n_1/(n_1+n_2) = \pi$. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. Coverage probabilities less than .925 or greater than .975 are in bold type.

transformation is often useful in reducing kurtosis in skewed and leptokurtic distributions. For example, taking the square root of Beta(1, 5) scores produced 95% coverage probabilities closer to .95 for all values of ρ_j. Although a confidence interval for $\mu_1 - \mu_2$ could be difficult to interpret with transformed data, data transformations do not introduce interpretation problems for the point-biserial correlation because it is a unitless measure of effect size.

The traditional method of constructing a confidence interval for η^2 uses a transformation of a computationally intensive confidence interval for an F non-centrality parameter (Steiger, 2004). The proposed confidence intervals for η^2 based on confidence intervals for ρ_1^2 and ρ_2^2 were examined using 100,000 Monte Carlo trials for random sample sizes (Table 11), fixed sample sizes (Table 12), normality, and homoscedasticity. The coverage probabilities of the 95% confidence intervals were close to .95 for all conditions except for $\rho_j = .2$, where the coverage probability was closer to .975.

Table 8. Estimated coverage probabilities for ρ_4 with normality and fixed sample sizes

			$\sigma_1/\sigma_2 = 1$		$\sigma_1/\sigma_2 = 2$	
ρ_4	n_1	n_2	CI for ρ_2	CI for ρ_4	CI for ρ_2	CI for ρ_4
0	15	15	.948	.955	.955	.955
	20	40	.950	.952	**.887**	.951
	40	20	.950	.952	**.982**	.954
	30	30	.949	.953	.947	.952
.2	15	15	.948	.956	.944	.954
	20	40	.950	.951	**.887**	.951
	40	20	.950	.951	**.981**	.954
	30	30	.949	.953	.947	.952
.4	15	15	.949	.957	.943	.954
	20	40	.951	.952	**.877**	.951
	40	20	.951	.952	.974	.953
	30	30	.949	.953	.945	.952
.6	15	15	.950	.958	.940	.954
	20	40	.953	.953	**.861**	.949
	40	20	.950	.953	.960	.953
	30	30	.950	.954	.940	.952
.8	15	15	.951	.958	.932	.952
	20	40	.957	.953	**.833**	.949
	40	20	.957	.953	**.919**	.952
	30	30	.950	.954	.931	.951

Note. $\rho_1 = \rho_4$ with $\sigma_1/\sigma_2 = 1$. $\hat{\rho}_2$ is not a consistent estimator of ρ_4 with $\sigma_1/\sigma_2 \neq 1$ and unequal fixed sample sizes. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. Coverage probabilities less than .925 or greater than .975 are in bold type.

The performance of the varying-coefficient confidence interval was compared with the constant-coefficient and random-coefficient methods for $m = 5$. The formulas given in Borenstein *et al.* (2009) were used to compute the constant-coefficient and random-coefficient confidence intervals. Four different patterns of ρ_2 were used in the computer simulations comparing the varying-coefficient and constant-coefficient confidence intervals. For each pattern, the varying-coefficient and constant-coefficient point estimates and confidence intervals were computed in 100,000 Monte Carlo trials. Within each trial, scores were randomly generated from a normal distribution with equal variances within and across the $m = 5$ studies. Table 13 compares the performance of the constant-coefficient confidence interval with the varying-coefficient confidence interval for fixed sample sizes of $n_1 = 20$, $n_2 = 30$, $n_3 = 40$, $n_4 = 50$, $n_5 = 60$, where n_j is the sample size per group within each of the $m = 5$ studies. The first row of Table 13 summarizes the performance of the two methods for equal values of ρ_2 across studies and thus satisfies a primary assumption of the constant,coefficient method. With effect-size equality, both the varying-coefficient and constant-coefficient methods yield nearly unbiased estimates of the average point-biserial correlation and both methods have 95% coverage probabilities that are close to .95. However, with effect-size heterogeneity,

Table 9. Estimated coverage probabilities for ρ_1 and ρ_3 with non-normality, homoscedasticity, and random sample sizes (n = 60)

		π = .20		π = .50	
Distribution	ρ_i	CI for ρ_1	CI for ρ_3	CI for ρ_1	CI for ρ_3
Beta(1, 1)	0	.952	.936	.949	.948
	.2	.951	.940	.950	.949
	.4	.950	.951	.955	.954
	.6	.949	.962	.963	.963
	.8	.952	**.990**	**.979**	**.979**
Beta(2, 2)	0	.952	.937	.949	.953
	.2	.951	.940	.950	.953
	.4	.949	.950	.953	.957
	.6	.946	.966	.960	.963
	.8	.945	.986	.972	.974
Beta(2, 4)	0	.952	.936	.949	.948
	.2	.951	.941	.949	.949
	.4	.947	.951	.951	.951
	.6	.941	.965	.955	.954
	.8	.936	**.981**	.960	.960
Beta(1, 5)	0	.953	.928	.950	.949
	.2	.949	.935	.948	.948
	.4	.940	.943	.942	.942
	.6	.924	.952	.934	.934
	.8	**.904**	.955	**.917**	**.917**
$\sqrt{\text{Beta}(1,5)}$	0	.952	.936	.949	.949
	.2	.950	.941	.949	.949
	.4	.948	.951	.951	.951
	.6	.943	.965	.956	.956
	.8	.939	**.982**	.964	.964

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated Beta (a, b) scores within each group. Coverage probabilities less than .925 or greater than .975 are in bold type. The random sample sizes (n_1 and n_2) were constrained to be >2.

which is the rule rather than the exception in practice, the varying-coefficient method continues to perform properly, while the performance of the constant-coefficient method is unacceptable with 95% coverage probabilities that are substantially <.95.

The serious limitations of the constant-coefficient meta-analysis methods are now well known (Schmidt & Hunter, 2015, p. 368). Bonett (2008a) derived an expression for the large-sample bias of the constant-coefficient point estimator which shows that this estimator is consistent with effect-size homogeneity or with equal weights. In practice, the weights that are used in a constant-coefficient meta-analysis will be unequal and the coverage probability of a 95% constant-coefficient confidence interval can be far less than .95, as illustrated in Table 13.

Random-coefficient methods, which do not assume effect-size homogeneity, have been proposed as a preferred alternative to constant-coefficient methods. Table 14

Table 10. Estimated coverage probabilities for ρ_2 and ρ_4 with non-normality, homoscedasticity, and fixed sample sizes

		$n_j = 15$		$n_j = 30$	
Distribution	ρ_i	CI for ρ_2	CI for ρ_4	CI for ρ_2	CI for ρ_4
Beta(1, 1)	0	.947	.955	.949	.952
	.2	.949	.956	.950	.950
	.4	.954	.961	.955	.958
	.6	.962	.968	.964	.967
	.8	**.978**	**.982**	**.979**	**.981**
Beta(2, 2)	0	.947	.955	.949	.953
	.2	.949	.957	.950	.953
	.4	.952	.959	.953	.957
	.6	.959	.965	.960	.963
	.8	.971	.975	.972	.974
Beta(2, 4)	0	.948	.955	.949	.952
	.2	.949	.957	.950	.953
	.4	.951	.958	.951	.955
	.6	.954	.961	.954	.958
	.8	.959	.966	.960	.963
Beta(1, 5)	0	.949	.958	.949	.953
	.2	.948	.957	.948	.953
	.4	.943	.953	.943	.948
	.6	.933	.945	.934	.940
	.8	**.916**	.929	**.915**	**.923**
$\sqrt{\text{Beta}(1, 5)}$	0	.948	.955	.948	.953
	.2	.948	.956	.950	.953
	.4	.951	.958	.952	.955
	.6	.955	.962	.956	.959
	.8	.964	.969	.964	.966

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated Beta (a, b) scores within each group. Coverage probabilities less than .925 or greater than .975 are in bold type.

summarizes the performance of the varying-coefficient and random-coefficient methods under some conditions that are nearly ideal for the random-coefficient methods. Within each of the 100,000 Monte Carlo trials, the ρ_2 values were randomly selected from a beta distribution of ρ_2 values, the sample sizes were randomly generated to minimize the correlation between the Fisher-transformed estimates and the weights, and the y-scores within each group were randomly generated from a normal distribution with equal variances within and across studies. The Beta(2, 2), Beta(3, 3), and Beta(4, 4) distributions are symmetric and unimodal. The Beta(3,3) and Beta(4, 4) distributions are also bell-shaped. The results in Table 14 show that the random-coefficient estimator of the average point-biserial correlation is biased and the 95% random-coefficient confidence interval has a coverage probability that is substantially less than .95. In contrast, the varying-coefficient estimator of the average point-biserial correlation is nearly unbiased and the 95% varying-

Table 11. Estimated coverage probabilities for ρ_1^2 with normality, homoscedasticity, and random sample sizes

	$\pi = .20$		$\pi = .50$	
ρ_1	$n = 30$	$n = 60$	$n = 30$	$n = 60$
0	.956	.952	.949	.949
.2	**.978**	**.976**	.972	.973
.4	.954	.945	.953	.949
.6	.945	.937	.949	.950
.8	.943	.929	.952	.950

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. The random sample sizes (n_1 and n_2) were constrained to be greater than 2. Coverage probabilities greater than .975 are in bold type.

Table 12. Estimated coverage probabilities for ρ_2^2 with normality, homoscedasticity, and fixed sample sizes

ρ_2	$n_1 = 15, n_2 = 15$	$n_1 = 20, n_2 = 40$	$n_1 = 30, n_2 = 30$
0	.948	.949	.949
.2	.973	.975	.974
.4	.951	.951	.949
.6	.949	.953	.949
.8	.951	.957	.950

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group.

Table 13. Comparison of constant-coefficient (CC) and varying-coefficient (VC) methods for ρ_2 with normality and homoscedasticity

		VC method		CC method	
ρ_2	ρ	Average estimate	95% coverage probability	Average estimate	95% coverage probability
[.3 .3 .3 .3 .3]	.3	.298	.947	.303	.944
[.1 .2 .3 .4 .5]	.3	.298	.946	.364	**.662**
[.5 .4 .3 .2 .1]	.3	.298	.949	.264	**.869**
[.5 .2 .1 .2 .5]	.3	.298	.948	.322	**.904**

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. The sample sizes per group were fixed at $n_1 = 20$, $n_2 = 30$, $n_3 = 40$, $n_4 = 50$, $n_5 = 60$ within each trial. Coverage probabilities less than .925 are in bold type.

coefficient confidence interval has coverage probabilities close to .95 under all conditions. Note that the greater bias in the random-coefficient estimator is due primarily to reverse-transforming an average of Fisher-transformed correlations with greater superpopulation heterogeneity.

When using a random-coefficient method, it is important to also report a confidence interval for the variance of the random effect. However, the currently available confidence intervals for the random-effect variance are hypersensitive to minor violations of the

Table 14. Comparison of varying-coefficient (VC) and random-coefficient (RC) methods for ρ_2 with normality and homoscedasticity

		VC method		RC method	
Distribution of ρ_2	ρ	Average estimate	95% coverage probability	Average estimate	95% coverage probability
Beta(4, 4) – .2	.3	.298	.947	.317	**.882**
Beta(3, 3) – .2	.3	.298	.947	.322	**.856**
Beta(2, 2) – .2	.3	.298	.947	.331	**.800**
Beta(2, 4.65)	.3	.299	.948	.322	**.862**

Note. Estimates in each row are based on 100,000 Monte Carlo trials using randomly generated normal scores within each group. The equal sample size per group for each study was randomly generated from a Uniform(20, 60) distribution within each trial. Coverage probabilities less than .925 are in bold type.

superpopulation normality assumption, and a very large number of studies are required to assess this critical assumption. A large number of studies are also needed to obtain a usefully narrow confidence interval for the random-effect variance. In contrast, pairwise comparisons or subgroup analyses can be used to effectively describe the nature of effect-size heterogeneity with varying-coefficient methods.

10. Hypothesis tests

10.1. Directional two-sided hypothesis test

In some applications, the researcher simply needs to decide if ρ_i is either greater than or less than some researcher-specified value (h). The sign of h will depend on how X is coded. If ρ_i is determined to be $> h$, this could provide support for one theory or one course of action, and if ρ_i is determined to be $< h$, then this could provide support for another theory or another course of action. This type of decision for a given value of α is called a directional two-sided test (see Jones & Tukey, 2000). It can be shown that the probability of making a directional error (i.e., deciding that $\rho_i > h$ when $\rho_i < h$ or deciding that $\rho_i < h$ when $\rho_i > h$) is at most $\alpha/2$, assuming all assumptions of the test have been satisfied.

A confidence interval for ρ_i can be used to conduct a directional two-sided test for the population point-biserial correlation. Specifically, if the lower limit of the $100(1 - \alpha)\%$ confidence interval is $> h$, then the null hypothesis ($\rho_i = h$) is rejected and we accept $\rho_i > h$; if the upper limit of the $100(1 - \alpha)\%$ confidence interval is less than h, then the null hypothesis is rejected and we accept $\rho_i < h$; if the $100(1 - \alpha)\%$ confidence interval includes h, the results are inconclusive.

For the special case of $h = 0$, the independent-samples t-test can be used to conduct a directional two-sided test for a population point-biserial correlation. Specifically, if the p-value is $< \alpha$ and the t-value is positive then accept $\rho_i > h$; if the p-value is $< \alpha$ and the t-value is negative then accept $\rho_i < h$. If the p-value is $> \alpha$, the results are inconclusive. For $h = 0$, the independent-samples t-test can be used for all four of the point-biserial correlations because $\mu_1 = \mu_2$ implies $\rho_i = 0$ ($i = 1, \ldots, 4$).

A confidence interval for $\rho_{i1} - \rho_{i2}$ can be used to conduct a directional two-sided test for a difference in two population point-biserial correlations that have been estimated from two independent samples. Specifically, if the lower limit of the $100(1 - \alpha)\%$ confidence

interval is > h, the null hypothesis ($\rho_{i1} - \rho_{i2} = h$) is rejected and we accept $\rho_{i1} - \rho_{i2} > h$; if the upper limit of the 100(1 – α)% confidence interval is less than h, the null hypothesis is rejected and we accept $\rho_{i1} - \rho_{i2} < h$; if the 100(1 – α)% confidence interval includes h, the results are inconclusive.

10.2. Equivalence test

In some two-group studies, the researcher wants to show that two different treatments (e.g., an inexpensive new treatment and the current treatment) or two different demographic subpopulations (e.g., men and women) have similar population means (Wellek, 2010). Suppose two treatments or two subpopulations are considered to be equivalent if $\mu_1 - \mu_2$ is within the $-h$ to h range, which is called the range of practical equivalence (ROPE). A confidence interval for $\mu_1 - \mu_2$ can be used to decide if $\mu_1 - \mu_2$ is inside or outside the ROPE. In applications where the response variable has an arbitrary metric or if the values of the response variable do not have clear clinical interpretations, it could be difficult for the researcher to specify a ROPE for $\mu_1 - \mu_2$. In these situations, it might be easier for the researcher to specify a ROPE for ρ_i. For example, a researcher might argue that a point-biserial correlation within the range −.1 to .1 represents a small or unimportant difference in population means. A 100(1 – 2α)% confidence interval for ρ_i can be used to decide if ρ_i is within the range $-h$ to h, or if ρ_i is outside this ROPE (Wellek, 2010). If the confidence interval for ρ_i is completely within the ROPE, the two treatments or subpopulations are declared to be equivalent; if the confidence interval for ρ_i is completely outside the ROPE, the two treatments or subpopulations are declared to be non-equivalent; and if the confidence interval includes the value $-h$ or h, the results are inconclusive.

10.3. Non-inferiority test

A 100(1 – α)% confidence interval for ρ_i can be used to conduct a non-inferiority test (Wellek, 2010). Suppose the ROPE is $-h$ to h and the goal of the study is to determine if $\rho_i > -h$ (non-inferiority) or $\rho_i < -h$ (inferiority). For this test, accept $\rho_i > -h$ if the lower limit for ρ_i is greater than $-h$, and accept $\rho_i < -h$ if the upper limit for ρ_i is less than $-h$. The results are inconclusive if the confidence interval includes the value $-h$ or h. In some applications, it is sufficient to show that an inexpensive treatment is not inferior to a more expensive treatment. The traditional *t*-test of equal population means is not an appropriate test for non-inferiority.

10.4. Directional non-equivalence test

A 100(1 – α)% confidence interval for ρ_i can be used to conduct other non-traditional hypothesis tests. For example, suppose the ROPE is $-h$ to h and the goal of the study is to determine if $\rho_i > h$ or $\rho_i < -h$. For this test, accept $\rho_i > h$ if the lower limit for ρ_i is greater than h, accept $\rho_i < -h$ if the upper limit for ρ_i is less than $-h$, and accept the hypothesis of equivalence if the confidence interval for ρ_i is completely within the $-h$ to h range. The results are inconclusive if the confidence interval includes the value $-h$ or h. Compared to the traditional test of $\rho_i = 0$, where a rejection of the null hypothesis does not preclude the possibility that ρ_i is very close to 0, the acceptance of $\rho_i > h$ or $\rho_i < -h$ indicates that the value of ρ_i is at least meaningfully large in addition to specifying the direction of the effect.

11. Sample size planning

11.1. Sample size for desired power

The point-biserial correlations presented here are useful supplements to the independent-samples *t*-test. When planning a two-group experiment with $n_2/n_1 = r$ and approximately equal variances, the values of n_1 and n_2 required for an independent-samples *t*-test with power $1 - \beta$ and a Type I error rate of α are very accurately approximated by the formula

$$n_1 = \tilde{\sigma}^2(1 + 1/r)(z_{\alpha/2} + z_\beta)^2/(\tilde{\mu}_1 - \tilde{\mu}_2)^2 + z_{\alpha/2}^2/4, \tag{34}$$

and $n_2 = rn_1$, where $\tilde{\sigma}^2$ is a planning value of the average within-group variance, $\tilde{\mu}_1 - \tilde{\mu}_2$ is a planning value of the expected difference in population means, $z_{\alpha/2}$ is a two-tailed critical *z*-value, z_β is a one-tailed critical *z*-value, and the adjustment $z_{\alpha/2}^2/4$ is based on results given by Guenther (1981).

Researchers might have difficulty using equation (34) if they have difficulty specifying $\tilde{\sigma}^2$ or the value of $\tilde{\mu}_1 - \tilde{\mu}_2$. Given the relation between a point-biserial correlation and standardized mean difference, equation (34) can be expressed as

$$n_1 = \frac{\left[\frac{(1-\tilde{\rho}^2)(1+r)}{4r}\right]\left(z_{\alpha/2} + z_\beta\right)^2}{\tilde{\rho}^2} + z_{\alpha/2}^2/4 \tag{35}$$

where $\tilde{\rho}$ is a planning value of ρ_2. A planning value of ρ_2 could be obtained from expert opinion, a pilot study, or a review of the literature. Some researchers will find equation (35) easier to implement than equation (34). As can be seen from equation (35), a smaller value of $\tilde{\rho}$ produces a larger sample size requirement. The R function `size.test.pbcor2` (Appendix) computes equation (35).

Equations (34) and (35) are appropriate for experimental designs. In a non-experimental design with simple random sampling, the total sample size ($n = n_1 + n_2$) required to conduct a directional two-sided test of H_0: $\rho_1 = 0$ with power $1 - \beta$ and a Type I error rate α is approximately

$$n = \frac{\left(1 - 1.5\tilde{\rho}^2 + \frac{\tilde{\rho}^2}{4\tilde{\pi}(1-\tilde{\pi})}\right)\left(z_{\alpha/2} + z_\beta\right)^2}{\tilde{\rho}^{*2}}, \tag{36}$$

where $\tilde{\pi}$ is a planning value of π and $\tilde{\rho}^{*2} = \ln[(1 + \tilde{\rho})/(1-\tilde{\rho})]/2$. The R function `size.test.pbcor1` (Appendix) computes equation (36).

11.2. Sample size for desired precision

The hypothesis testing result of an independent-samples *t*-test does not provide effect-size information. A confidence interval for a population point-biserial correlation will provide useful information about the magnitude of the effect if the confidence interval is sufficiently narrow. When planning a two-group experiment with $n_2/n_1 = r$ and approximately equal variances, the values of n_1 and n_2 required to obtain a $100(1 - \alpha)\%$ confidence interval for ρ_2 that has a desired width of about w are approximately

$$n_1 = \left(\frac{1+r}{r}\right)\left[\frac{\tilde{\rho}^2}{2(1-\tilde{\rho}^2)}+1\right]\left[\frac{\tilde{\rho}^2}{1-\tilde{\rho}^2}+1\right]^{-3}\left(\frac{z_{\alpha/2}}{w}\right)^2 \tag{37}$$

and $n_2 = rn_1$. The R function `size.ci.pbcor2` (Appendix) computes equation (37).

When planning a non-experimental study with simple random sampling, the total sample size required to obtain a $100(1-\alpha)\%$ confidence interval for ρ_1 that has a desired width of about w is approximately

$$n = \left[4\left(1-\tilde{\rho}^2\right)^2\left(1-1.5\tilde{\rho}^2+\frac{\tilde{\rho}^2}{4\tilde{\pi}(1-\tilde{\pi})}\right)\right]\left(\frac{z_{\alpha/2}}{w}\right)^2, \tag{38}$$

where $\tilde{\pi}$ is a planning value of π. The R function `size.ci.pbcor1` (Appendix) computes equation (38).

12. Examples

12.1. Example 1

Howell (2007, p. 200) described a two-group experiment to assess the effect of stereotype threat on mathematics examination performance of college students. The estimated means were 9.64 and 6.58, the estimated standard deviations were 3.17 and 3.03, and the sample sizes were 11 and 12 for the control and stereotype threat groups, respectively. Howell computed a pooled-variance independent-samples *t*-test and obtained $t(21) = 2.37$, $p = .027$. This result allows us to reject the null hypothesis of equal population means at $\alpha = .05$ and conclude that the population mean in the control condition is greater than the population mean in the stereotype threat condition.

To describe the magnitude of the population effect size, a confidence interval for $\mu_1 - \mu_2$, a standardized mean difference, or a point-biserial correlation should be reported along with the *t*-test result. The sample sizes are too small to assess homoscedasticity, and it is prudent to report a 95% confidence interval for ρ_4 rather than ρ_2. Using the R function `ci.pbcor124` (Appendix), the point estimate of ρ_4 is .446 and the 95% confidence interval for ρ_4 is [.037, .689]. It could be argued that this confidence interval is too wide to provide useful scientific or practical information and the study should be replicated using a larger sample size. Using the R function `size.ci.pbcor2` (Appendix) with $\alpha = .05$, a point-biserial planning value of .446, assuming equal sample sizes per group, and a desired 95% confidence interval width of .3, the required sample size per group in a replication study is about 50.

The sample size required to achieve desired precision is often substantially larger than the sample size required to achieve desired power of a two-sided directional test. Using the R function `size.test.pbcor2` (Appendix), the sample size required to conduct an independent-samples *t*-test with $\alpha = .05$, power of .9, and a point-biserial effect size of .446 is about 23 per group.

12.2. Example 2

Wright, Quick, Hannah, and Hargrove (2017) developed a new scale to measure 'character' in which one of the subscales was a measure of 'valour'. One of the goals of this study was to develop a new measure of valour that is not gender-biased (T. A. Wright, personal communication). They conducted two studies. In one study they obtained a

simple random sample of college students, and in a second study they obtained a simple random sample of working adults. The members of each sample were classified into male and female groups. Using the reported descriptive statistics in Wright *et al.* (2017) and utilizing the R function `ci.pbcor124` (Appendix), the estimate of ρ_1 is .052 with a 95% confidence interval of [–.053, .156] for the college students and –.068 with a 95% confidence interval of [–.187, .053] for the working adults. Using the R function `ci.diff.pbcor` (Appendix), a 95% confidence for the difference in ρ_1 values in the two populations is [–.040, .278]. This confidence interval includes 0, which could justify an examination of the average point-biserial correlation in the two populations. Using the R function `ci.ave.pbcor` (Appendix), an estimate of the average of the ρ_1 values in the two populations is –.008 with a 95% confidence interval of [–.088, .071]. Note that the confidence interval for the average point-biserial correlation is substantially narrower than the confidence interval for each separate population and narrow enough to perform an equivalence test. If we assume that a point-biserial correlation between valour and gender of less than about .1 is evidence of gender equivalence, then the confidence interval for the average point-biserial correlation suggests that the gender bias in the new valour scale is small and unimportant.

13. Conclusion

Each point-biserial correlation is appropriate in specific types of applications. The classical point-biserial correlation (ρ_1) is appropriate in both experimental and non-experimental designs if the homoscedasticity assumption can be justified. The classical point-biserial correlation also is appropriate in non-experimental designs with heteroscedasticity if simple random sampling is used. The ρ_2 measure of point-biserial correlation is appropriate in experimental designs with equal or unequal sample sizes if the homoscedasticity assumption can be justified. The ρ_3 measure of point-biserial correlation is appropriate in non-experimental designs with stratified random sampling, equal or unequal sample sizes, and heteroscedasticity. The ρ_4 measure of point-biserial correlation is appropriate in experimental designs with equal or unequal sample sizes and heteroscedasticity. The appropriate types of applications for the four point-biserial correlations are summarized in Table 15.

The current practice of reporting the *p*-value for an independent-samples *t*-test along with only a sample value of a standardized mean difference or a point-biserial correlation

Table 15. Summary of point-biserial formulas and applications

Parameter	Point estimator equation	Confidence interval equations	Applications
ρ_1	(12)	(25) and (26)	Experimental or non-experimental designs with equal or unequal n_j and equal σ_j^2; or non-experimental designs with unequal σ_j^2 and simple random sampling
ρ_2	(14)	(25) and (28)	Experimental designs with equal or unequal n_j and equal σ_j^2
ρ_3	(15)	(25) and (27)	Non-experimental designs with stratified random sampling, equal or unequal n_j and equal or unequal σ_j^2
ρ_4	(16)	(25) and (28)	Experimental designs with equal or unequal n_j and equal or unequal σ_j^2

can be more misleading than reporting only the *p*-value. In Example 1, the sample point-biserial correlation was .442 (which could interpreted as a 'large' effect), and it would be tempting to conclude that stereotype threat had not only a statistically significant effect but also a large effect on performance. However, the 95% confidence interval [.039, .687] for the population point-biserial correlation provides important additional information and suggests that the population point-biserial correlation could be trivial or very large. In this case, a larger sample is needed to more accurately assess the size of the stereotype threat effect. The confidence intervals for point-biserial correlations presented here can be used to supplement the results of an independent-samples *t*-test with useful effect-size information.

In studies with two independent samples, the *t*-test for equal population means is typically performed, but several non-traditional hypothesis tests (e.g., equivalence test, non-inferiority test, directional non-equivalence test) can also be performed. These non-traditional tests require the researcher to specify a ROPE, but this might be difficult to do in terms of a mean difference. When the effect size is expressed as a point-biserial correlation, it is usually easier to specify a ROPE. The confidence intervals for a population point-biserial correlation presented here can be used to perform a variety of useful non-traditional hypotheses tests.

The point-biserial correlation is a commonly used measure of effect size in meta-analyses. The currently used constant-coefficient and random-coefficient meta-analysis methods for point-biserial correlations have serious limitations, and their continued use is difficult to justify. The varying-coefficient meta-analysis methods for point-biserial correlation presented here have excellent performance characteristics and do not make any of the unrealistic assumptions of the constant-coefficient and random-coefficient methods. With the new point-biserial correlations introduced here, the most appropriate type of point-biserial correlation can be computed for each study and then combined in the meta-analysis.

In a study that has used a sample size that is too small, hypothesis tests will have low power and confidence intervals could be uselessly wide. Sample size planning is perhaps one of the most important steps in the design of a proposed study. The sample size formulas presented here can be used to design a study that will have an acceptably narrow point-biserial confidence interval or a hypothesis test with desired power.

References

American Psychological Association (2010). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Arnold, S. F. (1990). *Mathematical statistics*. Englewood Cliff, NJ: Prentice-Hall.

Bonett, D. G. (2008a). Confidence intervals for standardized linear contrasts of means. *Psychological Methods*, *13*, 99–109. https://doi.org/10.1037/1082-989X.13.2.99

Bonett, D. G. (2008b). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13*(3), 173–181. https://doi.org/10.1037/a0012868

Bonett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods*, *14*, 1–5. https://doi.org/10.1037/a0014270

Bonett, D. G., & Price, R. M. (2015). Varying coefficient meta-analysis methods for odds ratios and risk ratios. *Psychological Methods*, *20*, 394–406. https://doi.org/10.1037/met0000032

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior*, *28*, 647–659. org/10.1002/job.448

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York, NY: Harcourt Brace Jovanovich. https://doi.org/10.1037/1082-989X.9.2.164

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks, CA: Sage.

Goldberger, A. S. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.

Gradstein, M. (1986). Maximal correlation between normal and dichotomous variables. *Journal of Educational Statistics*, *11*, 259–261. https://doi.org/10.3102/10769986011004259.

Guenther, W. C. (1981). Sample size formulas for normal theory *T* tests. *American Statistician*, *35*, 243–244. org/10.1080/00031305.1981.10479363

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. https://doi.org/10.2307/1164588

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414. https://doi.org/10.1037/1082-989X.6.1.17

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*, 51–69. https://doi.org/10.1177/0013164404264850

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *4*, 386–401. https://doi.org/10.1037/1082-989X.11.4.386

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classified in a 2×2 table. *Biometrika*, *34*, 139–167. https://doi.org/10.1093/biomet/34.1-2.139

Scheffé, H. (1959). *The analysis of variance*. New York, NY: Wiley.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182. https://doi.org/10.1037/1082-989X.9.2.164

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? Hillsdale, NJ: Erlbaum.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics* (Vol. *2A*). London, UK: Arnold.

Tate, R. F. (1954). The correlation between a discrete and a continuous variable. *Point-biserial correlation. Annals of Mathematical Statistics*, *25*, 603–607. https://doi.org/10.1214/aoms/1177728730

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: CRC Press.

Wright, T. A., Quick, J. C., Hannah, S. T., & Hargrove, M. B. (2017). Best practice recommendations for scale construction in organizational research: The development and initial validation of the Character Strength Inventory (CSI). *Journal of Organization Behavior*, *38*, 615–628. org/10.1002/job.2180

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Received 13 February 2019; revised version received 21 August 2019

Appendix:

R functions

```
ci.pbcor124 <- function(alpha, m1, m2, sd1, sd2, n1, n2) {
 # Computes confidence intervals for three types of population
 # point-biserial correlations in a two-group design.
 # Arguments:
 #   alpha: alpha value for 1-alpha confidence
 #   m1:    sample mean for group 1
 #   m2:    sample mean for group 2
 #   sd1:   sample standard deviation for group 1
 #   sd2:   sample standard deviation for group 2
 #   n1:    sample size for group 1
 #   n2:    sample size for group 2
 # Returns:
 #   point estimate, SE, and CI for point-biserial correlations
 z <- qnorm(1 - alpha/2)
 n <- n1 + n2
 p <- n1/n
 b <- (n - 2)/(n*p*(1 - p))
 df1 <- n1 - 1
 df2 <- n2 - 1
 s1 <- sqrt((df1*sd1^2 + df2*sd2^2)/(df1 + df2))
 d1 <- (m1 - m2)/s1
 sed1 <- sqrt(d1^2*(1/df1 + 1/df2)/8 + 1/n1 + 1/n2)
 se1 <- sqrt((b^2*sed1^2)/(d1^2 + b)^3)
 se2 <- sqrt((16*sed1^2)/(d1^2 + 4)^3)
 lld1 <- d1 - z*sed1
 uld1 <- d1 + z*sed1
 cor1 <- d1/sqrt(d1^2 + b)
 cor2 <- d1/sqrt(d1^2 + 4)
 ll1 <- lld1/sqrt(lld1^2 + b)
 ul1 <- uld1/sqrt(uld1^2 + b)
 ll2 <- lld1/sqrt(lld1^2 + 4)
 ul2 <- uld1/sqrt(uld1^2 + 4)
 s2 <- sqrt((sd1^2 + sd2^2)/2)
 d2 <- (m1 - m2)/s2
 a1 <- d2^2*(sd1^4/df1 + sd1^4/df2)/(8*s2^4)
 a2 <- sd1^2/(s2^2*df1)+ sd2^2/(s2^2*df2)
 sed2 <- sqrt(a1 + a2)
 se4 <- sqrt((16*sed2^2)/(d2^2 + 4)^3)
 lld2 <- d2 - z*sed2
 uld2 <- d2 + z*sed2
 cor4 <- d2/sqrt(d2^2 + 4)
 ll4 <- lld2/sqrt(lld2^2 + 4)
 ul4 <- uld2/sqrt(uld2^2 + 4)
 out1 <- t(c(cor1, se1, ll1, ul1))
 out2 <- t(c(cor2, se2, ll2, ul2))
 out3 <- t(c(cor4, se4, ll4, ul4))
 out <- rbind(out1, out2, out3)
 colnames(out) <- c("Estimate", "SE", "LL", "UL")
 rownames(out) <- c("PB1", "PB2", "PB4")
 return(out)
}
```

Example

```
ci.pbcor124(.05, 9.64, 6.58, 3.17, 3.03, 11, 12)

     Estimate        SE         LL        UL
PB1 0.4588693 0.1629710 0.06095109 0.6969370
PB2 0.4428716 0.1601635 0.05830514 0.6808174
PB4 0.4424886 0.1678624 0.03719998 0.6886044
```

```
ci.pbcor3 <- function(alpha, m1, m2, sd1, sd2, n1, n2, p) {
 # Computes confidence intervals for a population point-biserial
 # correlation in a two-group nonexperimental design with stratified
 # random sampling.
 # Arguments:
 #   alpha: alpha value for 1-alpha confidence
 #   m1:    sample mean for group 1
 #   m2:    sample mean for group 2
 #   sd1:   sample standard deviation for group 1
 #   sd2:   sample standard deviation for group 2
 #   n1:    sample size for group 1
 #   n2:    sample size for group 2
 #   p:     proportion of subpopulation 1 members
 # Returns:
 #   point estimate, SE, and CI for point-biserial correlations
 z <- qnorm(1 - alpha/2)
 n <- n1 + n2
 b <- 1/(p*(1 - p))
 df1 <- n1 - 1
 df2 <- n2 - 1
 s2 <- sqrt(p*sd1^2 + (1 - p)*sd2^2)
 d2 <- (m1 - m2)/s2
 sed2 <- sqrt(d2^2*(1/df1 + 1/df2)/8 + (sd1^2/n1 + sd2^2/n2)/s2^2)
 se3 <- sqrt((b^2*sed2^2)/(d2^2 + b)^3)
 lld2 <- d2 - z*sed2
 uld2 <- d2 + z*sed2
 cor3 <- d2/sqrt(d2^2 + b)
 ll3 <- lld2/sqrt(lld2^2 + b)
 ul3 <- uld2/sqrt(uld2^2 + b)
 out <- t(c(cor3, se3, ll3, ul3))
 colnames(out) <- c("Estimate", "SE", "LL", "UL")
 rownames(out) <- c("PB3")
 return(out)
}
```

Example

```
ci.pbcor3(.05, 9.64, 6.58, 3.17, 3.03, 11, 12, .3)

       Estimate        SE         LL       UL
PB3   0.4151766 0.1548671 0.05315445 0.651824
```

```
ci.diff.pbcor <- function(alpha, cor1, ll1, ul1, cor2, ll2, ul2) {
 # Computes a confidence interval for a difference in two population
 # point-biserial correlations estimated from two different samples
 # Arguments:
 #   alpha: alpha value for 1-alpha confidence
 #   cor1:  sample point-biserial correlation in group 1
 #   ll1:   lower limit for first point-biserial correlation
 #   ul1:   upper limit for first point-biserial correlation
 #   cor2:  sample point-biserial correlation in group 2
 #   ll2:   lower limit for second point-biserial correlation
 #   ul2:   upper limit for second point-biserial correlation
 # Returns:
 #   confidence interval
 ll <- cor1 - cor2 - sqrt((cor1 - ll1)^2 + (ul2 - cor2)^2)
 ul <- cor1 - cor2 + sqrt((ul1 - cor1)^2 + (cor2 - ll2)^2)
 ci <- c(ll, ul)
 return(ci)
}
```

Example

```
ci.diff.pbcor(.05, .052, -.053, .155, -.069, -.186, .052)
[1] -0.03920612  0.27687816


ci.ave.pbcor <- function(alpha, cor, se) {
 # Computes confidence interval for an average point-biserial correlation
 # using estimates from two or more studies. Different types of point-biserial
 # correlations (PB1, PB2, PB3, PB4) can be used across studies.
 # Args:
 #   alpha: alpha level for 1-alpha confidence
 #   cor:   vector of point-biserial estimates
 #   se:    vector of point-biserial standard errors
 # Returns:
 #   estimated average, standard error, confidence interval
 m <- length(cor)
 z <- qnorm(1 - alpha/2)
 ave <- sum(cor)/m
 var.ave <- sum(se^2)/m^2
 cor.f <- log((1 + ave)/(1 - ave))/2
 ll0 <- cor.f - z*sqrt(var.ave/(1 - ave^2)^2)
 ul0 <- cor.f + z*sqrt(var.ave/(1 - ave^2)^2)
 ll <- (exp(2*ll0) - 1)/(exp(2*ll0) + 1)
 ul <- (exp(2*ul0) - 1)/(exp(2*ul0) + 1)
 out <- cbind(ave, sqrt(var.ave), ll, ul)
 colnames(out) <- c("Estimate", "SE", "LL", "UL")
 return(out)
}
```

Example

```
cor = c(.052, -.069)
se = c(.0533, .0613)
ci.ave.pbcor(.05, cor, se)

      Average         SE          LL         UL
[1,]  -0.0085 0.04061582 -0.08788419 0.07099147

size.test.pbcor2 <- function(alpha, cor, pow, r) {
 # Computes the sample size per group required to conduct a directional
 # two-sided test of a population point-biserial correlation with desired
 # power in a 2-condition experiment. Equality of variances is assumed.
 # Arguments:
 #   alpha:  alpha level for hypothesis test
 #   cor:    planning value of point-biserial correlation
 #   pow:    desired power
 #   r:      n2/n1 ratio
 # Returns:
 #   required sample size per group
 za <- qnorm(1 - alpha/2)
 zb <- qnorm(pow)
 k <- (1 - cor^2)*(1 + r)/(4*r)
 n1 <- ceiling(k*(za + zb)^2/(cor^2) + za^2/4)
 n2 <- n1*r
 out <- t(c(n1, n2))
 colnames(out) <- c("n1", "n2")
 return(out)
}
```

Example

```
size.test.pbcor2(.05, .446, .9, 1)
     n1 n2
[1,] 23 23
```

```
size.test.pbcor1 <- function(alpha, cor, pow, p) {
 # Computes the total sample size required to conduct a directional
 # two-sided test of a population point-biserial correlation with
 # desired power in a two-group nonexperimental design with simple
 # random sampling.
 # Arguments:
 #   alpha:  alpha level for hypothesis test
 #   cor:    planning value of point-biserial correlation
 #   pow:    desired power
 #   p:      proportion of subpopulation 1 members planning value
 # Returns:
 #   required sample size per group
 za <- qnorm(1 - alpha/2)
 zb <- qnorm(pow)
 cor.f <- log((1 + cor)/(1 - cor))/2
 k <- 1 - 1.5*cor^2 + cor^2/(4*p*(1 - p))
 out <- ceiling(k*(za + zb)^2/cor.f^2)
 return(out)
}
```

Example

```
size.test.pbcor1(.05, .3, .8, .25)
[1] 81
```

```
size.ci.pbcor2 <- function(alpha, cor, w, r) {
 # Computes the sample size required to estimate a population point-biserial
 # correlation with desired confidence and precision. Equality of variances
 # is assumed.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   cor:    planning value of point-biserial correlation
 #   w:      desired confidence interval width
 #   r:      n2/n1 ratio
 # Returns:
 #   required sample size per group
 z <- qnorm(1 - alpha/2)
 k1 <- (1 + r)/r
 k2 <- cor^2/(2*(1 - cor^2)^2) + 1
 k3 <- (cor^2/(1 - cor^2) + 1)^(-3)
 n1 <- ceiling(k1*k2*k3*(z/w)^2)
 n2 <- n1*r
 out <- t(c(n1, n2))
 colnames(out) <- c("n1", "n2")
 return(out)
}
```

Example

```
size.ci.pbcor2(.05, .446, .3, 1)
     n1 n2
[1,] 50 50
```

```
size.ci.pbcor1 <- function(alpha, cor, w, p) {
 # Computes the sample size required to estimate a population point-biserial
 # correlation with desired confidence and precision. Equality of variances
 # is assumed.
 # Arguments:
 #   alpha:  alpha level for 1-alpha confidence
 #   cor:    planning value of point-biserial correlation
 #   w:      desired confidence interval width
 #   p:      proportion of subpopulation 1 members planning value
 # Returns:
 #   required total sample size
 z <- qnorm(1 - alpha/2)
 out <- ceiling(4*((1 - cor^2)^2)*(1 - 1.5*cor^2 + cor^2/(4*p*(1 - p)))*(z/w)^2)
 return(out)
}
```

Example

```
size.ci.pbcor1(.05, .3, .2, .3)
[1] 310
```